DOCUMENT RESUME

ED 175 944 TM 009 686

AUTHOR Steelo, Joe M.
TITLE The Assessment of Writing Proficiency via Qualitative Ratings of Writing Samples.
PUB DATE 9 Apr 79
NOTE 10p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; *Composition Skills (Literary); *Essay Tests; Evaluators; Higher Education; Performance Tests; Persuasive Discourse; *Rating Scales; *Scoring; Testing; *Test Reliability; Test Validity; Writing Skills
IDENTIFIERS *Holistic Evaluation; *Interrater Reliability; Test Length

ABSTRACT

The College Outcome Measures Project/American College Testing Program (COMP/ACT) Writing Assessment is described, and issues of validity and reliability in the assessment of writing samples using qualitative rating scales are explored. COMP/ACT is composed of three role-playing tasks in the social sciences, natural sciences, and arts, which are recorded on audiotapes. Evaluators are chosen for their expertise in written communication. Qualitative rating scales are used to define five levels of proficiency on three broad criteria: the degree to which the writer addresses the intended audience; the development and organization of pertinent information; and the skillful use of language. Results of several studies indicated that these scales are sensitive to various levels of proficiency and can be applied in a highly reliable manner. Increasing the number of writing samples per student to three significantly increased reliability, but increasing the number of raters beyond two did not greatly enhance reliability. Use of a refined set of training materials for raters resulted in even higher levels of reliability. (The use of analytic as well as holistic rating procedures was also explored). (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED175944

THE ASSESSMENT OF WRITING PROFICIENCY VIA QUALITATIVE RATINGS OF WRITING SAMPLES

By

Joe M. Steele
College Outcome Measures Project
The American College Testing Program
P.O. Box 168
Iowa City, Iowa 52243

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Joe M. Steele

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Issues of validity and reliability in the assessment of writing samples using qualitative rating scales are explored. Results of several studies indicate that these scales are sensitive to various levels of proficiency and can be applied in a highly reliable manner. Increasing the number of writing samples per student to three significantly increases reliability, but increasing the number of raters beyond two does not greatly enhance reliability. Using a refined set of training materials for raters results in even higher levels of reliability. Use of analytic as well as holistic rating procedures are also explored.

PRINTED IN U.S.A.

April 1979

Presented at the Annual Meeting of the National Council on Measurement in Education in San Francisco, April 9, 1979.

TM009 686

THE ASSESSMENT OF WRITING PROFICIENCY VIA QUALITATIVE RATINGS OF WRITING SAMPLES

Joe M. Steele
The American College Testing Program

Writing is held to be an ability that is extremely important for effective functioning in a variety of adult roles, yet the evaluation of writing proficiency via writing samples can lead to disagreement among judges as to the levels of proficiency demonstrated. Such disagreement can result in unreliable evaluation of proficiency. Furthermore, writing tasks used in these assessments often bear little resemblance to real-life adult writing tasks. Hence, the validity of these measures as reflecting abilities important to effective functioning is also open to question.

Since 1976, the College Outcome Measures Project (COMP) of the American College Testing Program (ACT) has been developing college exit level measures that use realistic tasks to measure applied general education skills. The initial efforts by COMP to develop standardized, content-fair, and cost-effective methods of assessment of general education skills have been encouraging. An open-ended Measurement Battery was developed to assess Communicating, Solving Problems, Clarifying Values, Functioning within Social Institutions, Using Science and Technology, and Using the Arts. In two test-retest studies of this Battery, total test score reliabilities of .90 and .89 were achieved for the six-hour, open-ended examination.

One component of the COMP/ACT Assessment Materials is a measure of writing proficiency. This paper describes these COMP Writing Assessment procedures, addresses several aspects of validity of this assessment, and then reports results from reliability studies which examine:

1) the consistency with which this assessment procedure can be applied by faculty evaluators.
2) the degree to which improvements in evaluator training materials can increase interrater reliability.
3) the degree to which reliability can be influenced by varying the number of judges or the number of writing samples used.

Writing Assessment Procedures

The COMP/ACT Writing Assessment is composed of three role-playing tasks using audio-taped stimuli in the content areas of social sciences, natural sciences and the arts, which require a total of 60 minutes of writing to compose three written communications directed to various audiences. The rating procedure used to score these writing samples requires that a person expert in written communication (e.g. college-level English faculty) apply qualitative rating scales that define five levels of proficiency on three broad criteria to each of the three writing samples for each student. Guidelines are provided to minimize various sources of bias in rating.

For each task, respondents have twenty minutes in which to prepare their written communication to specified audiences. In Task 1, the respondent listens to a taped group discussion of "Residential Family" and then writes a personal letter to a relative that explains and reflects

on an invitation to join this "family." In Task 2, the respondent listens to a taped newscast on the energy shortage and then writes a persuasive letter to the editor of a local newspaper reacting to the newscast and suggesting a local response to the problem. In Task 3, the respondent listens to a tape of two types of music and then writes a memorandum to an employer endorsing one type as background music appropriate for the firm's waiting room.

The evaluation task calls for a holistic judgment of each writing sample of the degree to which the communicator addresses the intended audience, develops and organizes the points called for in the role-playing situation, and makes skillful use of language. The overriding guideline for the judgment is the degree to which the response represents effective writing for an adult engaged in this social role.

Equal emphasis is given to rhetorical concerns (organization), psychological concerns (making contact with and attending to the perspective of the audience), and practical concerns (using vivid language, illustrations, etc. to dramatize and create an effect). Little emphasis is given to formal errors such as misspelled words, punctuation, etc., unless they noticeably detract from or obscure the message. These criteria are qualitatively scaled, to allow a norm-free judgment of the overall effect of a piece of writing.

Faculty evaluators are also provided with the following guidelines to minimize various sources of bias:

1) All students should be rated on one writing task before the next set of responses is rated.
2) Penmanship should be disregarded and no credit or penalty be given for a position or attitude that you endorse or that you cannot accept.
3) When all students have been rated on a task, the first half dozen should be rerated to assure that interpretation of the scale did not shift after experience was developed in applying the scale.
4) Before beginning to rate the next writing sample, student responses should be stacked in a different order.
5) Avoid attempting to find out the identity, gender, age, grade level, or other characteristics of the students.

Validity Issues

From the beginning, the development of the COMP rating scale for writing samples has involved considerable input from faculty experts in communication at a diverse group of institutions participating in the College Outcome Measures Project. This process has assured that the outcomes as defined by the rating scale represent writing skill expectations which relevant faculty view as important and which are held in common by a diverse group of postsecondary institutions. Thus a high content validity for the scale has been established.

A second aspect of validity has to do with the nature and design of the writing tasks themselves. One feature which measures of writing have often lacked in the past has been a standardized task. For example, respondents have sometimes been asked to select a topic, such as "My summer vacation," from a list of topics. The context, other than a broad

general topic is entirely defined by the respondent. Such an approach introduces considerable variation into the nature of the response and makes comparisons of student performance difficult. Moreover, the task is not realistic. It does not deal with a situation an adult normally would respond to in writing, is not directed to any audience other than the evaluator, and does not set forth a purpose or outcome the communication should achieve.

COMP has taken a different approach. Rather than selecting a topic that draws on one's private fund of experience, the respondent is asked to play a specific role likely to be encountered by adults. Relevant details are provided as a context in which to respond. The individual is encouraged to use knowledge and skills acquired from any source, but the content needed to respond is information the generally educated college graduate might be expected to know. The COMP Writing Assessment is part of a Battery designed to measure college exit-level skills in general education. As students may be more knowledgeable in their own areas of specialization, the assessment is made content fair by providing a problem in each of the three broad areas of social sciences, natural sciences, and the arts. Thus, at least one of the writing tasks should provide an opportunity to demonstrate writing proficiency in a content area with which the respondent is familiar.

A third concern related to the validity of the assessment has been to assure that it was sensitive to the levels of writing proficiency acquired by undergraduate students through normal postsecondary activities. Note that this issue relates not only to the sensitivity of the instrument, but to the degree to which the outcomes measured are amenable to instruction. The combined mean rating (range 0-12) of seniors at seven institutions was found to be significantly higher than the mean for a comparable reference group of freshmen, as shown in Table 1. Thus, it would appear that the instrument is sensitive to differences in writing proficiency across groups who would be expected to differ.

A fourth validity concern was to establish that the writing skills, defined by the rating scale, are important to effective functioning in a variety of relevant adult roles. As part of a study of successful college graduates at the University of Nebraska-Lincoln (UNL), data were obtained on levels of writing proficiency (Forrest, 1979). Those graduates identified a priori as effective communicators obtained a mean score significantly higher than college seniors on the COMP writing assessment.

One might assume that a sample of college seniors contains some who are (or will be) effective in communication roles and some who are not. Based on this assumption, one would expect a group of experienced adults, selected because of their effective involvement in communication roles, to score higher than college seniors.

If the graduates had not scored higher than seniors, the data would have raised a question about the validity of this writing assessment as tapping a dimension important to effective functioning. While the evidence does not prove a necessary relationship, the data in Table 1 suggest that writing as assessed by COMP is important to effective functioning in a variety of adult roles related to communicating.

TABLE 1
A Cross-sectional Comparison of COMP Writing Performance
Across Levels of Educational Development

LEVEL	N	COMP SCORE Mean	COMP SCORE S.D	t-STATISTIC
College Freshmen	131	4.65	1.92	2.89*
College Seniors	96	5.39	1.73	9.15*
Effectively functioning Adult Communicators	42	7.69	2.26	

*$p < .001$

Reliability Issues

Study 1: (1) To what extent are the assessment procedures characterized by interrater agreement?
(2) What would be the effect of varying the number of items (writing samples) and raters?

Two studies of interrater reliability are reported in this paper. In the first study, three experienced communication faculty independently rated three writing samples per student for 61 freshmen.

The raters were provided with the stimulus materials, questions, and rating scales, along with some opportunity to discuss the scales and to try them out on some sample student responses. After rating sample responses for each task, raters discussed their ratings with each other and explained what cues they were attending to that resulted in discrepant ratings. Following this training session, the faculty evaluators independently completed the rating task.

The procedures used to estimate interrater reliability across three essays and three raters in the first study relied upon analyses of variance components due to respondents, raters, and writing tasks. More specifically, the procedures that were used are based upon generalizability theory as explicated by Cronbach and associates (1972) and partially summarized by Brennan (1977). The resulting generalizability coefficients reflect the degree of consistency across raters and can be interpreted as specific types of correlation coefficients. One assumption applied in deriving these coefficients is that both raters and writing tasks are random samples from an essentially infinite universe of possible raters and writing samples.

An analysis of the ratings of three essays by the three raters employed in study one yielded an estimated interrater reliability of .70, reflecting a great deal of consistency in ratings across raters.

In addition, generalizability coefficients were estimated for various numbers of writing tasks and raters, as reported in Table 2.

* Thanks are due to Robert Brennan for consultation concerning these analyses.

TABLE 2
Generalizability Coefficients (and error variances)*
for Study One Varying Number of Raters and Writing Samples

Number of Writing Samples	NUMBER OF RATERS 2	3
1	.43 (.483)	.49 (.317)
2	.58 (.269)	.62 (.224)
3	.65 (.197)	.70 (.162)
6	.75 (.125)	.79 (.099)
9	.79 (.101)	.83 (.078)

These results suggest that a moderately high degree of agreement was achieved among raters with regard to judgments of writing proficiency based on three writing samples, following the procedures developed for training raters. In addition, the results suggest that reducing the number of writing samples significantly reduces reliability, while increasing the number of samples beyond three or the number of raters beyond two does not seem to greatly enhance reliability.

Study 2: (1) Do refined evaluator training procedures enhance interrater reliability?
(2) Can an analytic rating procedure yield reliable part scores?

A second reliability study was conducted to study the effect of improvements in evaluator training materials and procedures on reliability of rating. In the second study, two experienced communication faculty independently rated three writing samples per student for 50 freshmen.

In addition to the guidelines provided in the first study, raters were given a written discussion of ways to identify various problems of rating and ways to compensate for or correct them. A set of ten sample student responses at many levels of proficiency were selected for each of the three writing tasks. Raters were asked to rate each sample response and then compare the rating to an agreed-upon standard rating for that sample. The two raters did not meet together but followed these training materials independently and after rating the sample responses each rated the fifty sets of writing samples. Table 3 shows the generalizability coefficients for Study 2.

* In Table 2 and subsequent tables, a number in parenthesis is an estimated error variance denoted $\hat{\sigma}^2(\delta)$ by Cronbach et al. (1972). This is the error variance for deviation scores, usually associated with comparisons among examinees. In interpretating this error variance, the reader should recall that the rating for a writing sample ranges from 0 to 4.

TABLE 3
Generalizability Coefficients (and error variances)
for Study two using raters, assuming items and raters
are random samples from an essentially infinite universe

Number of Writing Samples	Reliability for Two Raters
1	.58 (.403)
2	.73 (.201)
3	.76 (.144)

Table 3 is directly comparable to the estimates in Table 2. Improved training materials resulted in increasing the generalizability coefficient for 2 raters and 3 writing samples from .65 to .76. A pattern of increases in reliability for more raters or more writing samples similar to study 1 occurred in this study; however, only results for two raters and 1-3 writing samples are reported. Given the level of reliability obtained for two raters and three writing samples, the increased cost in student and evaluator time to write and score twice as many writing samples would not seem to be warranted for the relatively small increase in reliability.

The results reported in Table 3 address the generalizability of COMP/ACT writing scores assuming a replication using different raters and writing samples. If the analysis is done with writing samples fixed but raters assumed random, results address questions involved in a replication in which different raters rated the same writing samples. This condition speaks to the question of interrater agreement. Table 4 shows that under this assumption, a coefficient of .84 was obtained for the holistic rating. These results indicate that the COMP/ACT Writing Assessment can be applied in a highly reliable manner.

TABLE 4
Generalizability coefficients (and error variances)
reflecting reliability of ratings for two raters and three writing samples
varying rating criteria and universe score assumptions

SCORE	ASSUMPTION	
	Writing Samples and Raters Random	Writing Samples fixed and raters random
Holistic	.76 (.144)	.84 (.086)
Language subscore	.83 (.114)	.91 (.064)
Organization subscore	.74 (.161)	.86 (.085)
Audience subscore	.48 (.275)	.79(.112)
Total Analytic	.82 (.118)	.90 (.062)

Another reliability issue addressed by Table 4 has to do with whether a holistic rating or an analytic rating is made. An analytic rating procedure rates a list of writing characteristics rather than requiring a single holistic judgment. The rating scales for the COMP Writing Assessment could be used such that a separate rating would be given for each of the criteria (Language, Organization, Audience) rather than the holistic judgment that subjectively takes these criteria into account. An exploratory study was made of differences that would result from this rating convention. The same two evaluators used in the second study were asked to record analytic ratings as well as holistic ratings for the 50 sets of writing samples.

The Pearson product moment correlation of the total analytic rating with the holistic rating was found to be .97. This correlation is no doubt inflated as the same raters applied both rating procedures. This relationship does suggest that evaluators are taking the criteria into account in making the more cost-efficient holistic rating. Table 4 compares the generalizability coefficients for holistic and analytic scores for two raters and three writing samples under two assumptions. While this pilot study is limited, the reliability of both rating conventions appears to be quite high and error variance for both are quite similar.

Conclusions

The series of studies described in this paper provide strong encouragement that writing samples can be reliably rated. They suggest that measures such as the COMP/ACT Writing Assessment can provide a valid and objective way to measure writing proficiency using writing samples. Using such instruments, normative data can be generated for college freshmen and seniors. These data will be highly relevant in evaluating the results of the current public school emphasis in basic skills. They will provide assistance to colleges in setting appropriate standards for writing purposes.

Instruments such as the COMP/ACT Writing Assessment should make a significant contribution to the examination and improvement of college programs to develop writing proficiency.

References

Brennan, L. Generalizability Analyses: Principles and Procedures. ACT Technical Bulletin No. 26 Iowa City, Iowa: The American College Testing Program, September 1977.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. The Dependability of Behavorial Measurements: Theory of Generalizabilty for Scores and Profiles. New York: Wiley, 1972.

Forrest, A. A study of UNL College of Arts and Sciences Graduates. COMP/ACT Technical Report. Iowa City, Iowa: The American College Testing Program, February 1979.